Multiple Linear Regression

Yesoda Bhargava, MPH, University of York, UK.

Multiple Regression

- In multiple regression we use more than one explanatory variable to explain or predict a single response variable.
- The descriptive tools - scatterplots, least-squares regression, and correlation are essential preliminaries to inference and also provide a foundation for confidence intervals and significance tests.
- The introduction of several explanatory variables leads to many additional considerations.

Inference for Multiple Regression

Population multiple regression equation

- The simple linear regression model assumes that the mean of the responsible variable y depends on the explanatory variable x according to a linear equation

  $$\mu_y = \beta_0 + \beta_1 x$$

  .
- For any fixed value of x, the response y varies normally around this mean and has a standard deviation σ that is the same for all values of x.
- In the multiple regression the response variable y depends on not one but p explanatory variables. We will denote these explanatory variables by x_1, x_2, ..., x_p. The mean response is a linear function of the explanatory variables:

  $$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

  .
- This expression is the **population regression equation.**
- We can think of subpopulations of responses, each corresponding to a particular set of values for all of the explanatory variables $x_1, x_2, ..., x_p$.
- In each subpopulation, y varies normally with a mean given by the population regression equation. The regression model assumes that the standard deviation σ of the responses is the same in all subpopulations.

- Data used : Birth weight in MASS library of R.
- Response variable: Birth weight (BWT).
- Explanatory Variables : Age (AGE), Race (RACE), Weight at Last menstrual period (LWT).
- We will use these variables to predict the response variable BWT. There are three, n=3, explanatory variables. x_1=AGE, x_2=RACE and x_3= LWT.
- The birth weights of babies are given in grams.
- Data for a simple linear regression consists of obervations (x_i, y_i) on the two variables. Because there are several explanatory variables in multiple regression, the notation needed to describe the data is more elaborate.
- In our case study, each obervation or case consists of a value for the response variable and for each of the explanatory variables for each woman.
- Call x_{ij} the value of the jth explanatory variable for the ith woman.

- The data are then:

  Woman 1: $(x_{11}, x_{12}, x_{13}, ..., x_{1n}, y_1)$
  Woman 2: $(x_{21}, x_{22}, x_{23}, ..., x_{2n}, y_2)$

  Woman n: $(x_{p1}, x_{p2}, x_{p3}, ..., x_{np}, y_n)$

- Here, n is the number of cases and p is the number of explanatory variables. Data are often entered into computer regression programs in this format.
- Each row is a case and each column corresponds to a different variable.

Multiple Linear Regression Model

- We combine the population regression equation and assumptions about variation to construct the multiple linear regression model.
- The sub population means describe the FIT part of our statistical model. The RESIDUAL part represents the variation of observations about the means.
- The symbol ε represents the deviation of an individual observation from its subpopulation mean. We assume that these deviations are normally distributed with mean 0 and an unknown standard deviation σ that does not depend on the values of the x variables.
- This is an assumption that we can check by examining the residuals in the same way that we did for simple linear regression.
- The statistical model for multiple linear regression is
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$
  for i=1,2,....., n.
- The mean response μ_y is a linear function of the explanatory variables:
  $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.
- The deviations ε_i are independent and normally distributed with mean 0 and standard deviation σ. In other words, they are an SRS from the N(0,σ) distribution.
- The parameters of the model are β_0, β_1, β_2.....β_p and σ.

- The assumption that the subpopulation means are related to the regression coefficients β by the equation
  $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.
  implies that we can estimate all subpopulation means from estimates of the β's. To the extent that this is accurate, we have a useful tool for describing how the mean of y varies with the x's.
- For simple linear regression we used the principle of least squares to obtain the estimators of the intercept and slope of the regression line. For multiple regression the principle is the same but the details are more complicated.
- Let b_0, b_1, b_2, ..., b_p denote the estimators of the parameters β_0, β_1, β_2,.....β_p.
- For ith observation the predicted response is
  $\hat{y}_i = b_0 + b_1 x_{i1} + b_{i2} x_{i2} + ... + b_p x_{ip}$.
- The ith residual, the difference between the observed and the predicted response, is therefore
  $e_i =$ observed response-predicted response
  $= y_i - \hat{y}_i$
  $= y_i - b_0 - b_1 x_{i1} - b_{i2} x_{i2} - .... - b_p x_{ip}$
- The method of least squares chooses the values of b's that make the sum of the squares of the residuals as small as possible. In other words, the parameter estimates b_0, b_1, b_2, ..., b_p, minimize the quantity
  $\sum (y_i - b_0 - b_1 x_{i1} - b_{i2} x_{i2} - .... - b_p x_{ip})^2$

- The formula for the least-squares estimates is complicated.
- The parameter σ^2 measures the variability of the responses about the population regression equation. As in the case of simple linear regression, we estimate σ^2 by an average of the squared residuals. The estimator is
  $$s^2 = \frac{\sum e_i^2}{n-p-1}$$
  $$= \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$
- The quantity n-p-1 is the degrees of freedom associated with s^2. The degrees of freedom equals sample size n minus (p+1), the number of βs we must estimate to fit the model.
- In the simple linear regression case there is just one explanatory variable, so p=1 and the degrees of freedom are n-2. To estimate σ we use
  $s = \sqrt{s^2}$.

Confidence intervals and significance tests for regression coefficients

- We can obtain confidence intervals and significance tests for each of the regression coefficients β_j as we did in simple linear regression. The standard errors of the $b's$ have more complicated formulas, but all are multiples of s.
- We reply on statistical software for the calculations.
- A level C confidence interval for β_j is
  $b_j \pm t * SE_{b_j}$
  where SE_{b_j} is the standard error of b_j and t* is the value for the t(n-p-1) density curve with area C between -t* and t*.
- To test the hypothesis H_0: β_j=0, compute the t statistic
  $t = \frac{b_j}{SE_{b_j}}$.
- The test of hypothesis could be one-sided or two-sided.
- Because the regression is often used for prediction, we may wish to construct confidence intervals for a mean response and prediction intervals for a future observation from multiple regression models.

ANOVA table for multiple regression

- In simple linear regression the F test from the ANOVA table is equivalent to the two-sided t test of the hypothesis that the slope of the regression line is 0.
- For the multiple regression there is a corresponding ANOVA F-test, but it tests the hypothesis that all of the regression coefficients (with the exception of the intercept) are 0.
- Here is the general form of the ANOVA table for multiple regression.

Table 1: General form of the ANOVA table for multiple regression

	Df	Sum of squares	Mean Square	F Value
Model	p	SSM $=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	MSM= $\frac{SSM}{DFM}$	F= $\frac{MSM}{MSE}$
Error	n-p-1	SSE= $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	MSE= $\frac{SSE}{DFE}$	
Total	$n-1$	SST= $\sum_{i=1}^{n}(y_i-\bar{y})^2$	MST= $\frac{SST}{DFT}$	

- The ANOVA table is similar to that for the simple linear regression. The degrees of freedom for the model increase from 1 to p reflect the fact that now we have p explanatory variables rather than just one. As a consequence, the degrees of freedom for error decrease by the same amount.
- It is always a good idea to calculate the degrees of freedom by hand and then check that your software agrees with your calculations. In this way you can verify your software is using the number of cases and the number of explanatory variables you intended.

- The sums of squares represent sources of variation. Once again, both the sum of squares and their degrees of freedom add:
  SST=SSM + SSE
  DFT=DFM + DFE.
- The estimate of the variance σ^2 for our model is again given by the MSE in the ANOVA table. That is $s^2 =$ MSE.
- The ratio of MSM/MSE is an F statistic for testing the null hypothesis:
  $H_0 : \beta_1 = \beta_2 = \beta_3..... = \beta_p = 0$
  against the alternative hypothesis
  H_a :at least one of the β_j is not zero.
- The null hypothesis says that none of the explanatory variables are predictors of the response variable when used in the form expressed by the multiple regression equation.
- The alternative states that atleast one of them is linearly related to the response. As in simple linear regression, large values of F gives evidence against H_0.
- When H_0 is true, F has the F(p,n-p-1) distribution. The degrees of the freedom for the F distribution are those associated with the model and error in the ANOVA table.
- *A common error in the use of multiple regression is to assume that all of the regression coefficients are statistically different from zero whenever the F statistic has a small P-value. Be sure that you understand the difference between the F test and t tests for individual coefficients.*

Squared multiple correlation R^2

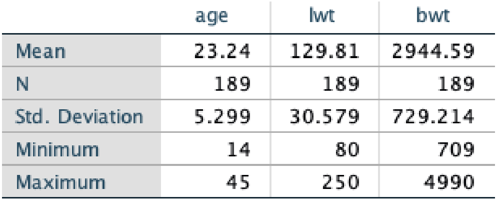
- For simple linear regression we notes that the square of the sample correlation could be written as the ratio of SSM to SST and could be interpreted as the proportion of variation in y explained by x.
- A similar statistic is routinely calculated for multiple regression. The statistic

  $$R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

  is the proportion of the variation of the response variable y that is explained by the explanatory variables $x_1, x_2, x_3, ..., x_p$ in a multiple linear regression.
- Often R^2 is multiplied by 100 and expressed as a percent. The square root of R^2, called the multiple correlation coefficient is the correlation between the observations y_i and the predicted values $\hat{y}_i$.

Case study: Preliminary analysis

- In this case study we analyse the data from birth weight dataset. The response variable is the birth weight. The explanatory variables as previously mentioned are LWT, AGE and RACE.
- We also examine smoking status (SMOK) and hypertension (HTN) as explanatory variables. Total size of the data is 189. We use SPSS to illustrate the outputs.
- The first step in the analysis is to carefully examine each of the variables.



	age	lwt	bwt
Mean	23.24	129.81	2944.59
N	189	189	189
Std. Deviation	5.299	30.579	729.214
Minimum	14	80	709
Maximum	45	250	4990

Figure 1: Descriptive statistics for the AGE, LWT and BWT in the dataset.

Interpretation

- The maximum value of AGE variables appears to be rather extreme: it is (45-23.24)/5.299 = 4.1 standard deviations above the mean. We do not discard this case at this time but will take care in our subsequent analyses to see if it has an excessive influence on our results.
- The maximum value of LWT variable appears to be rather extreme: it is (250-129.81)/30.579 = 3.93 standard deviations above the mean. We do not discard this case at this time but will take care in our subsequent analyses to see if it has an excessive influence in our results.
- The mean birth weight of baby is 2.9kg and the minimum is 0.7 kg, more than 2.5 standard deviations away from the mean.
- We look at the distribution of these variables using histogram.
- AGE and LWT appears right-skewed and BWT shows a slight left skew. Note that the outlier in the AGE distribution is the 45 years which we mentioned above is 4.1 standard deviations away from the mean.
- **It is important to note that the multiple regression model does not require any of these distributions to be normal.**
- The purpose of examining these plots is to understand something about each variable alone before attempting to use it in a complicated model.
- **Extreme value of each variable should be noted and checked for accuracy.** If found to be correct, the cases with these values should be carefully examined to see if they are truly exceptional and perhaps do not belong in the same analysis with the other cases.

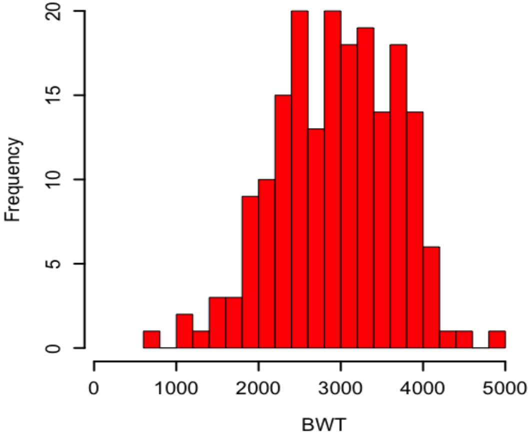


Figure 2: Distribution of BWT in the dataset

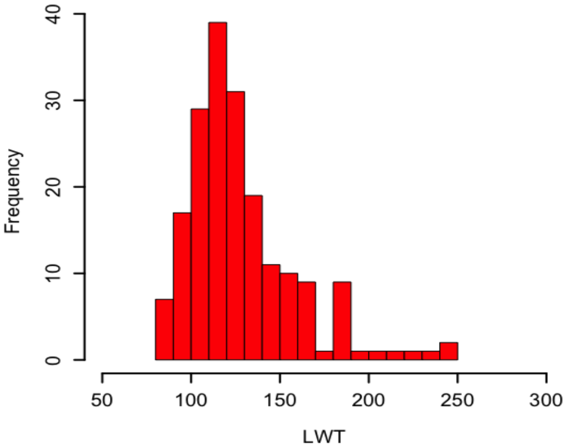


Figure 3: Distribution of LWT in the dataset

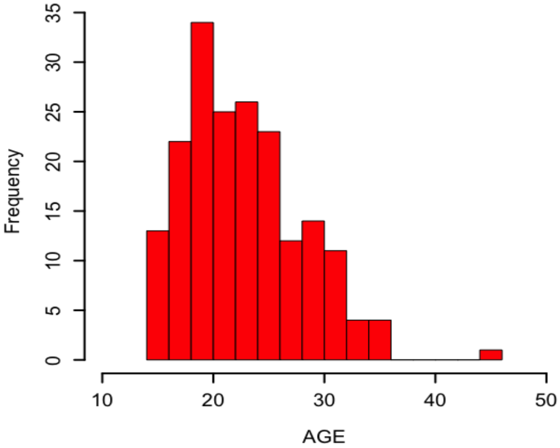
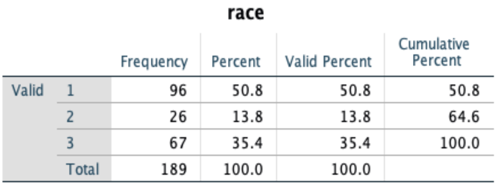


Figure 4: Distribution of AGE in the dataset

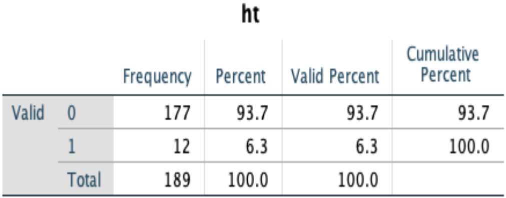


race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	96	50.8	50.8	50.8
	2	26	13.8	13.8	64.6
	3	67	35.4	35.4	100.0
	Total	189	100.0	100.0	

Figure 5: Descriptive statistics for RACE variable in the dataset.

We see that close to 51% of the mothers in the sample belong to white ethnicity and 35% to other ethnicity, remaining 14% to Black. White ethnicity people form the majority in the group (96 out of 189).

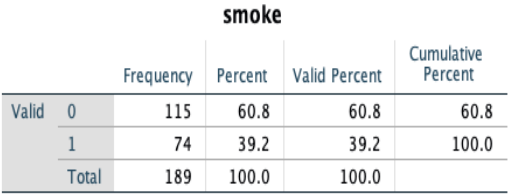


ht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	177	93.7	93.7	93.7
	1	12	6.3	6.3	100.0
	Total	189	100.0	100.0	

Figure 6: Descriptive statistics for HTN variable in the dataset.

From the tabular summary of hypertension in the data, we observe that 94% have reported absence of any diagnosis of hypertension while remaining (12 out of 189) have reported the condition. Majorly the women do not have the health condition in this dataset.



smoke

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	115	60.8	60.8	60.8
	1	74	39.2	39.2	100.0
	Total	189	100.0	100.0	

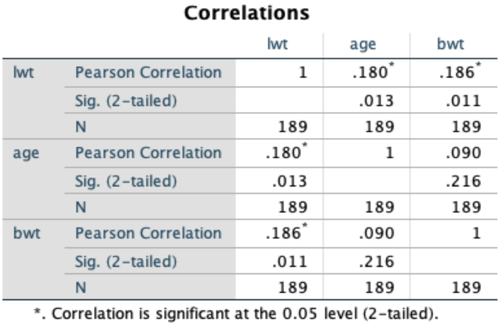
Figure 7: Descriptive statistics for SMOK variable in the dataset.

For smoking data, we find that almost 39% of the women in the dataset have reported smoking. This appears to be quite a higher prevalence of smoking.

Relationships between pairs of variables : Continuous variables

The second step in our analysis is to examine the relationships between all pairs of variables. Scatter plots and correlations are our tools for studying two-variable relationships.

From the Fig. 8 , we see that the correlation between LWT and AGE is 0.180 wih a p-value of 0.013, and with BWT it is 0.186 with a p-value of 0.011 . Both are statistically significant by any reasonable standard. The correlation between AGE and BWT is 0.090 with a p-value of 0.216 indicating that the correlation is not found to be statistically significant, however, the direction reported is positive.



Correlations

		lwt	age	bwt
lwt	Pearson Correlation	1	.180*	.186*
	Sig. (2-tailed)		.013	.011
	N	189	189	189
age	Pearson Correlation	.180*	1	.090
	Sig. (2-tailed)	.013		.216
	N	189	189	189
bwt	Pearson Correlation	.186*	.090	1
	Sig. (2-tailed)	.011	.216	
	N	189	189	189

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 8: Correlation between AGE, BWT and LWT. p value is reported

Continuous vs Categorical Variable

- Note that the variables HTN, RACE and SMOK are categorical in nature so we will use side-by-side box plots to assess the relationship between continuous variables (AGE, BWT and LWT) and these variables.
- Note that the range of age is high for non-hypertensive woman while those who have reported hypertension have ages roughly between 16 to 25 years. The variation og age is less in woman with hypertension, mostly young women have reported hypertension.

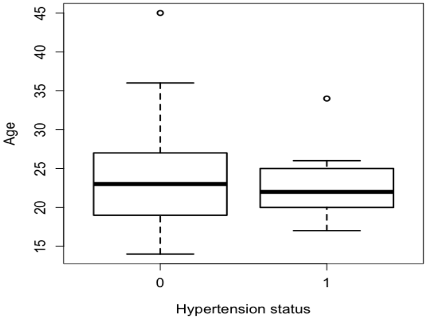


Figure 9: Relationship between AGE and Hypertension in the data. (1 indicates Yes for Hypertension and 0 indicates a No.)

Continuous vs Categorical Variable

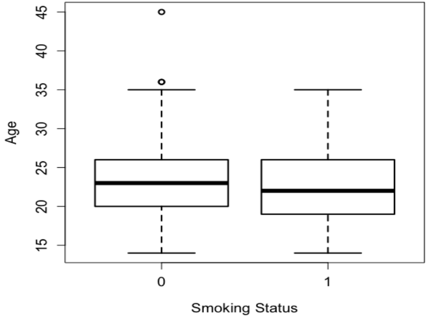


Figure 10: Relationship between AGE and Smoking in the data. (1 indicates Yes for smoking, 0 indicates No.)

Not vast differences between the distribution of age is observed for smoking and non-smoking women.

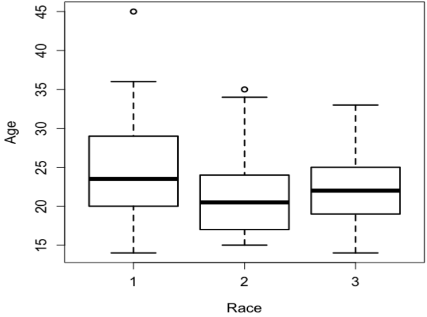


Figure 11: Relationship between AGE and Race in the data. (1 indicates White, 2 indicates Black and 3 indicates Other.)

Age distribution for white has a higher inter-quartile range, a right skew is possible in age for Black group , Other groups show fairly evenly-balanced distribution of age.

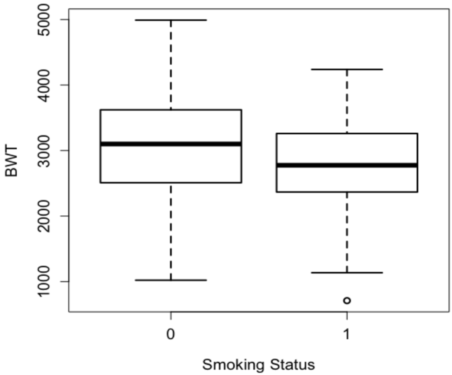


Figure 12: Relationship between BWT and Smoking in the data. (1 indicates Yes for smoking, 0 indicates No.)

Median birth weight of babies born to women who smoke is found to be slightly lower than that of women who do not smoke. The highest weight of babies born to women who smoke is also less than those babies born to non-smoking mothers. There does appear to be a difference in birth weight by smoking status of the mother.

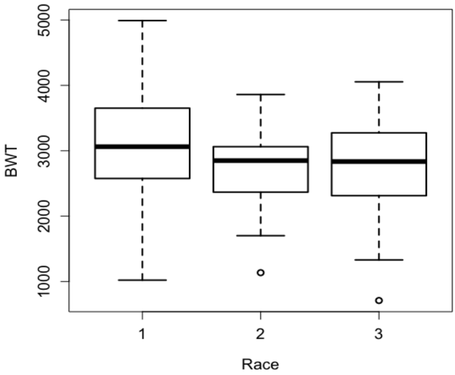


Figure 13: Relationship between BWT and Race in the data. (1 indicates White, 2 indicates Black and 3 indicates Other.)

Median birth weight is found to be almost similar for all three ethnic groups. A wide range is observed for Whites but not so for other ethnicities. It perhaps, reflects the smaller size of these ethnicities (26 Black and 67 Others)

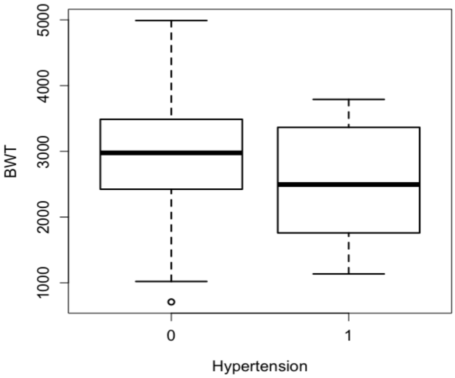


Figure 14: Relationship between BWT and Hypertension in the data. (1 indicates Yes and 0 indicates N for hypertension.)

- Median birth weight of babies born to women who have hypertension is less than birth weight of babies born to non-hypertensive women. The inter-quartile range of birth weight is seen to be higher for the former category and the highest birth weight is lower than the highest birth weight of kids born to women without hypertension. The plot hints at some likely relationship between hypertension and birth weight of babies.

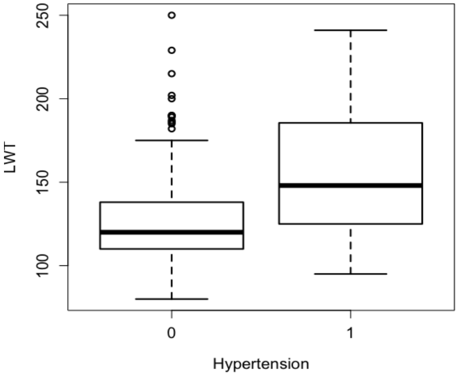


Figure 15: Relationship between LWT and Hypertension in the data. (1 indicates Yes and 0 indicates N for hypertension.)

- Weight of women who have no reported hypertension before last menstrual period seems to be lesser in magnitude as compared to that of women who have reported hypertension. Minimum, median and maximum value are all lower for the former group as compared to the latter. Several outliers are also seen in the non-hypertensive women indicating a right skew.

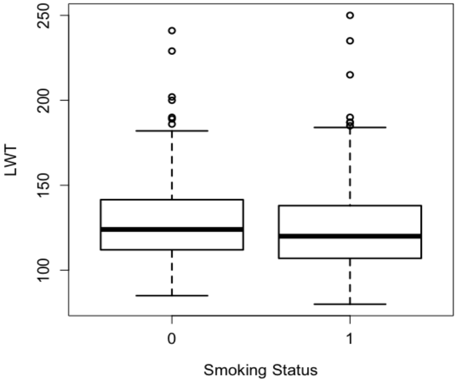


Figure 16: Relationship between LWT and Smoking in the data. (1 indicates Yes for smoking, 0 indicates No.)

No outright differences are observed for the last menstrual weight of two groups of women separated by smoking status. The distribution appears quite similar.

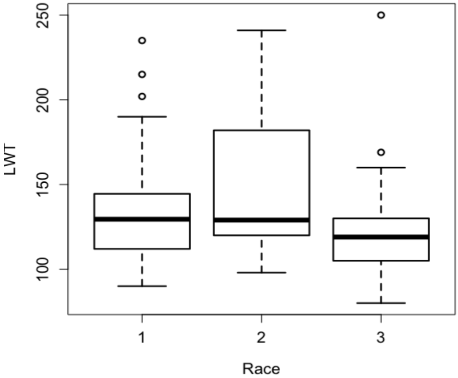
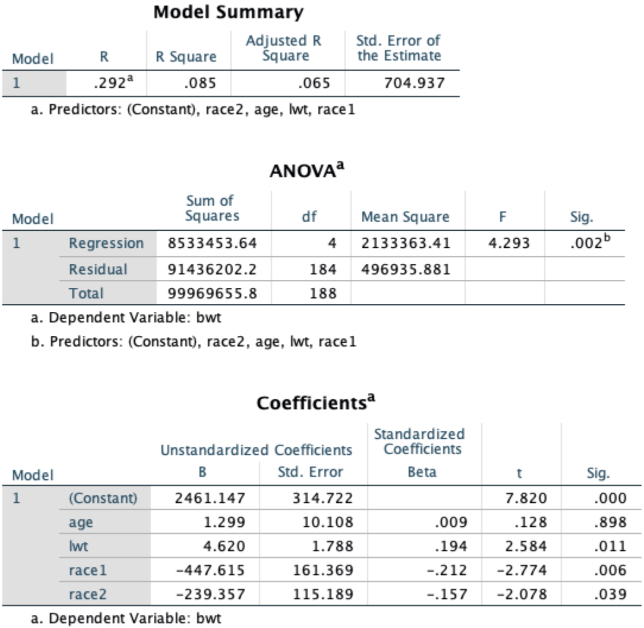


Figure 17: Relationship between LWT and Race in the data. (1 indicates White, 2 indicates Black and 3 indicates Other.)

The distribution of mother's weight just before pregnancy are different across the race categories. The women belonging to Black appear to be heavier than the other two groups, as indicated by wide variability. The highest weight is found for women belonging to Black ethnic category and lowest for the Other group. This group also shows an outlier at 250 pounds.

Regression on birth weight

- To explore the relationship between the explanatory variables and our response variable, BWT, we run several multiple regressions. We divide the variables into two groups for deeper understanding of the inclusion of variables in the regression model: Age, LWT, RACE and later we add SMOK and HTN.
- We need to create dummy variable for RACE because it is a categorical variable with more than 2 levels. We use race1 and race2 such that race1=1 when its black and race2=1 when the woman belongs to other group. When both race1 and race2 are zero, it refers to the White group.
- We begin our analysis by using AGE, LWT and RACE to predict BWT.
- Fig. 18 gives the multiple regression output.
- The output contains an ANOVA table, some additional descriptive statistics, and information about the parameter estimates.
- When examining the ANOVA table it is a good idea to first verify the degrees of freedom. This ensures that we have not made some serious error in specifying the model for the program or in entering the data. There are a total of 189 cases, we have DFT = 189-1=188. There are four explanatory variables so DFM = 4 and DFE = 188-4=184. Note that dummy variable for race results in one extra variable, so instead of RACE we use RACE1 and RACE2, and hence the total number of explanatory variables becomes 4.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.292[a]	.085	.065	704.937

a. Predictors: (Constant), race2, age, lwt, race1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8533453.64	4	2133363.41	4.293	.002[b]
	Residual	91436202.2	184	496935.881		
	Total	99969655.8	188			

a. Dependent Variable: bwt

b. Predictors: (Constant), race2, age, lwt, race1

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2461.147	314.722		7.820	.000
	age	1.299	10.108	.009	.128	.898
	lwt	4.620	1.788	.194	2.584	.011
	race1	-447.615	161.369	-.212	-2.774	.006
	race2	-239.357	115.189	-.157	-2.078	.039

a. Dependent Variable: bwt

Figure 18: Multiple regression output for regression using age, lwt and race to predict bwt.

- The ANOVA F statistic is 4.293 with a p-value of 0.002. Under the null hypothesis:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  the F statistic has an F(4,184) distribution. According to this distribution, the chance of obtaining an F statistic of 4.293 or larger is 0.002. We therefore conclude that atleast one of the four regression coefficients for the birth weight is different from 0 in the population regression equation.
- In the Model Summary, we find that the MSE is 496935.881, thus the root MSE is 704.93, this value is s, the estimate of the parameter σ of our model. The value of R^2 is 0.085. That is, 8.5% of the observed variation in the birth weight is explained by linear regression on age, lwt and race.
- Although the p-value is very small (0.002) the model does not explain very much of the variation in birth weight. **Remember, a small P-value does not necessarily tell us that we have a large effect, particularly when the sample size is large.**
- From the parameter estimates we obtain the **fitted regression equation**
  $B\hat{W}T = 2461.147 + 1.299AGE + 4.620LWT - 447.6RACE1 - 239.357RACE2.$
- Lets find the predicted BWT for a woman of Black ethnicity aged 30 years with LWT of 145 pounds. The explanatory variables are AGE = 30, LWT=145, RACE1=1, RACE2=0. The predicted BWT is:
  $$\begin{aligned} B\hat{W}T &= 2461.147 + 1.299(30) + 4.620(145) - 447.6(1) - 239.357(0) \\ &= 2722.417\ . \end{aligned}$$

- Recall that the t statistics for testing the regression coefficients are obtained by dividing the estimates by their standard errors. Thus, for the coefficient of AGE we obtain the t-value given in the output by calculating
  $t = \frac{b}{SE_b} = \frac{1.299}{10.108} = 0.1285$
  .
- The p-values appear in the last column. Note that these p-values are for the two-sides alternatives. AGE has a p-value of 0.898 , and we conclude that the regression coefficient for this variable is not significantly different from 0 whereas for LWT it is (p-value = 0.011). The P-values for RACE1 (0.006)and RACE2 (0.039) also achieve statistical significance.
- Interpretation: We would say that for every one unit increase in LWT given all other values remain the same, the birth weight of the baby increases by 4.62gm and that the birth weight of a baby born in Black ethnic group has 447.6 gm weight less on average than that of the baby born to a White mother. Furthermore, the birth weight of Other category baby is found to be 239.35 gms less on average than that of the baby born to White mother.

Residuals

- As in simple linear regression, we should always examine the residuals as an aid to determining whether the multiple regression model is appropriate for the data.
- Because there are several explanatory variables, we must examine several residual plots. It is usual to plot the residuals versus the predicted values $\hat{y}$ and also versus each of the explanatory variables.
- Look for outliers, influential observations, evidence of a curved rather than linear relation, and anything else unusual.
- If the deviations ε in the model are normally distributed, the residuals should be normally distributed. Fig. 19 presents a normal quantile plot of the residuals. The distribution appears to be approximately normal.

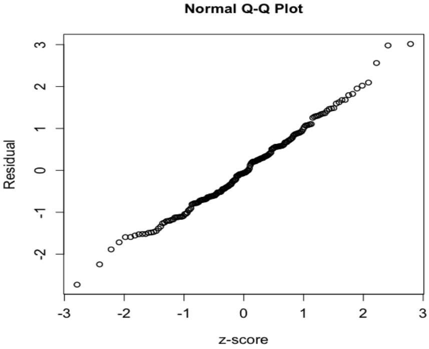


Figure 19: Normal quantile plot of the residuals from the birth weight model. There are no important deviations from normality.

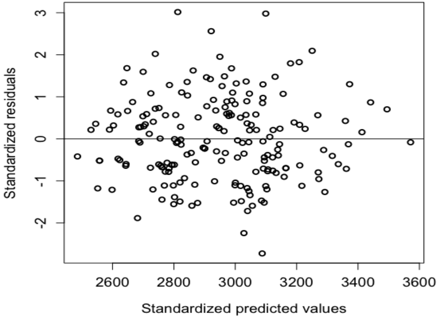
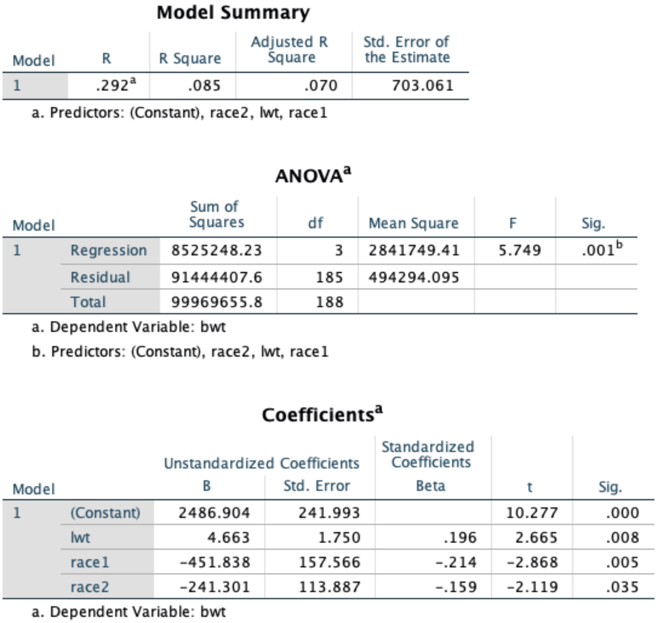


Figure 20: Scatter plot of standardized residual versus standardized predicted values for the birth weight model. The residuals show no pattern in distribution and are roughly homogeneously distributed around the regression line (residual =0)

Refining the model

- Because the variable AGE has the largest P-value of the four explanatory variables (See Fig. 18) and therefore appears to contribute the least to our explanation of birth weight, we rerun the regression using only LWT and RACE. The output appears in Fig. 21.
- The F statistic indicates that we can reject the null hypothesis that the regression coefficients for the three explanatory variables is zero. The P-value has reduced and is now 0.001. The value of R^2 is the same at 8.5%. Dropping of AGE from the model resulted in the loss of very little explanatory power, in fact F statistic improved (5.749).
- The measure s of variation for the new model is 703.06 which is nearly identical for the two regressions, another indication that dropping AGE loses very little.
- The t statistics for the individual regression coefficients indicate that LWT is still clearly significant (p=0.008) while the coefficient of both RACE1 and RACE2 are slightly larger than before in absolute value terms and approach the statistical significance (0.005 and 0.035 respectively).



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.292[a]	.085	.070	703.061

a. Predictors: (Constant), race2, lwt, race1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8525248.23	3	2841749.41	5.749	.001[b]
	Residual	91444407.6	185	494294.095		
	Total	99969655.8	188			

a. Dependent Variable: bwt

b. Predictors: (Constant), race2, lwt, race1

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2486.904	241.993		10.277	.000
	lwt	4.663	1.750	.196	2.665	.008
	race1	-451.838	157.566	-.214	-2.868	.005
	race2	-241.301	113.887	-.159	-2.119	.035

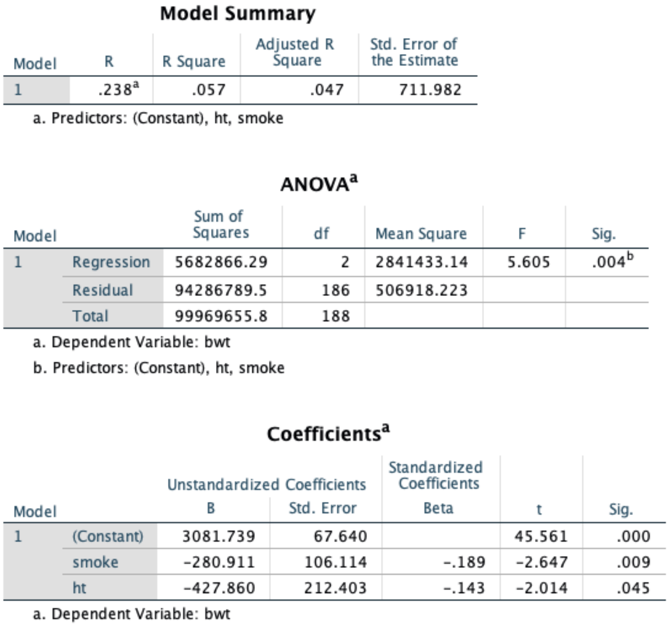
a. Dependent Variable: bwt

Figure 21: Multiple regression output for regression using lwt and race to predict bwt.

- Comparison of the fitted equations for the two multiple regression analyses tells us something more about the intricacies of thi procedure.
- For the first run we have $B\hat{W}T = 2461.147 + 1.299AGE + 4.620LWT - 447.6RACE1 - 239.357RACE2.$ whereas the second gives us $B\hat{W}T = 2486.904 + 4.663LWT - 451.838RACE1 - 241.301RACE2.$
- Eliminating AGE from the model changes the regression coefficients for all of the remaining variables and the intercept. This phenomenon occurs quite generally in multiple regression.
- **Individual regression coefficients, their standard errors, and significance tests are meaningful only when interpreted in the context of the other explanatory variables in the model.**

Regression on Smoking and Hypertension status

- Note that both smoking and hypertension are binary variables so we do not need to create dummy variables. Dummy variables need only be created when the categorical variable has more than two levels.
- We now turn to the problem of predicting BWT using the two variables SMOK and HTN. Fig. 22 gives the output. The fitted model is $B\hat{W}T = 3081.739 - 280.911SMOK - 427.860HTN$.
- The degrees of freedom are as expected: 2, 186, 188. The F statistic is 5.605 with a p-value of 0.004. We conclude that the regression coefficients for SMOK and HTN are not both zero. Recall that we obtained a p-value of 0.001 when we used LWT and RACE to predict BWT. Both multiple regression equations are highly significant, but this obscures the fact that the two models have quite different explanatory power.
- For this regression R^2 is 5.7% whereas for the previous regression it was 8.5%, a higher value (Fig. 21).
- Stating that we have a statistically significant result is quite different from saying that an effect is large or important.
- Further examination of the output in Fig. 22 reveals that the coefficient for both SMOK (p=0.009)and HTN (p=0.045) are statistically significant.
- For a complete analysis we should carefully examine the residuals.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.238[a]	.057	.047	711.982

a. Predictors: (Constant), ht, smoke

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5682866.29	2	2841433.14	5.605	.004[b]
	Residual	94286789.5	186	506918.223		
	Total	99969655.8	188			

a. Dependent Variable: bwt

b. Predictors: (Constant), ht, smoke

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3081.739	67.640		45.561	.000
	smoke	-280.911	106.114	-.189	-2.647	.009
	ht	-427.860	212.403	-.143	-2.014	.045

a. Dependent Variable: bwt

Figure 22: Multiple regression output for regression using SMOK and HTN to predict BWT.

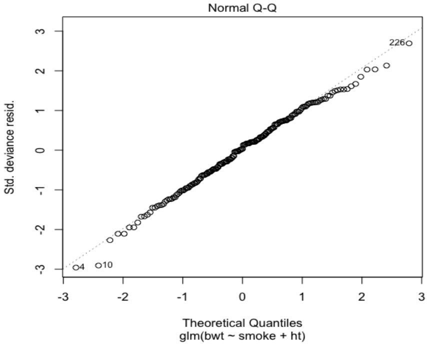


Figure 23: Normal quantile plot of the residuals from the smoking and hypertension model. There are no important deviations from normality. We can see 3 outliers but they fall relatively closer to the line.

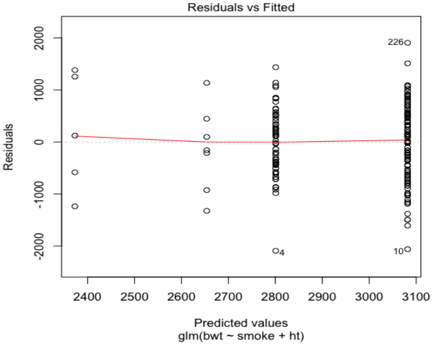
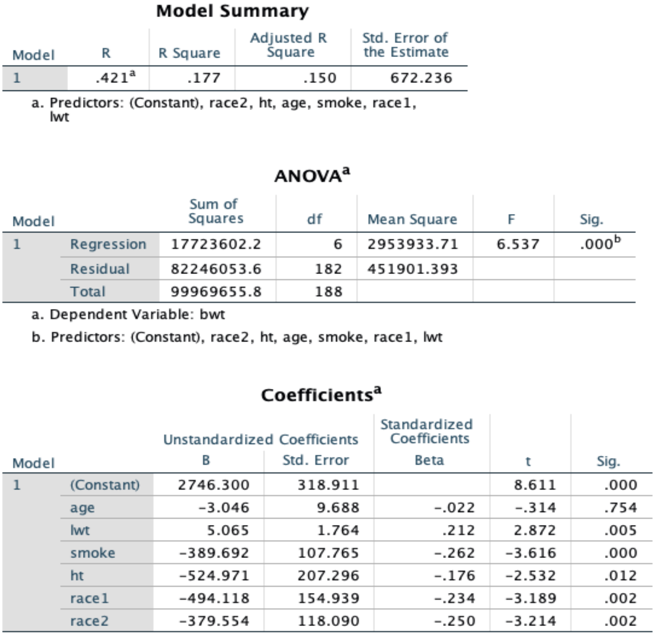


Figure 24: Scatter plot of residual versus predicted values for the birth weight model.

There are four groups we can see. Because both the predictors are binary, there will be four groups of outcomes and therefore we see the plot like this. Note that for all the group of predicted values, the residuals are distributed fairly uniformly around the residual=0 line, hence the uniformity of variance appears to be satisfied.

Regression using all variables

- We have seen that LWT and RACE and SMOK, HTN give a highly significant regression equation.Comparing the values of R^2 for the two models indicates that LWT and RACE are better predictors than SMOK and HTN. Can we get a better prediction equation using all of the explanatory variables together in one multiple regression?
- To address this question we run the regression with all 4 explanatory variables. The output appears in Fig. 25.
- Note the degrees of freedom, DFT=6, DFE = 188-6 = 182.
- The F statistic is 6.537 with a p-value less than 0.0001, so atleast one of our four explanatory variables has a non-zero regression coefficient in the population regression line. This result is not surprising given that we have already seen that LWT, RACE, SMOK and HTN are strong predictors of BWT.
- The value of R^2 is 17.7% which is impressively high from 5.7% which indicates that the explanatory power of the model has increased. The root MSE is 672.2 which is less than the previous values obtained indicating that the variation in the residuals have decreased which is evident in a higher value of R^2 and F statistic.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.421[a]	.177	.150	672.236

a. Predictors: (Constant), race2, ht, age, smoke, race1, lwt

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17723602.2	6	2953933.71	6.537	.000[b]
	Residual	82246053.6	182	451901.393		
	Total	99969655.8	188			

a. Dependent Variable: bwt

b. Predictors: (Constant), race2, ht, age, smoke, race1, lwt

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2746.300	318.911		8.611	.000
	age	-3.046	9.688	-.022	-.314	.754
	lwt	5.065	1.764	.212	2.872	.005
	smoke	-389.692	107.765	-.262	-3.616	.000
	ht	-524.971	207.296	-.176	-2.532	.012
	race1	-494.118	154.939	-.234	-3.189	.002
	race2	-379.554	118.090	-.250	-3.214	.002

Figure 25: Multiple regression output for regression using all variables to predict BWT.

- Examination of the t statistics and the associated P-values for the individual regression coefficients reveals that AGE is the only variable which is not significant (p=0.754) while all others provide evidence for statistical significance.
- Smoking (-389.692, p <0.0001) and hypertension (-524.971, p=0.012) are negatively associated with the birth weight meaning that women who smoke or are hypertensive have babies born with a less body weight as compared to women who do not smoke or are not hypertensive.
- LWT is positively associated (5.065, p =0.005) with birth weight and for every unit increase in it, the average birth weight of the baby is found to increase by 5 gms, given all other parameters are kept constant.
- Both races Black (-494.118, p=0.002)and Other (-379.554, p = .002) have a significantly less average of birth babies as compared to White for all common value of other parameters.
- After every regression we must examine the residual plots to verify if our model is suited for prediction.

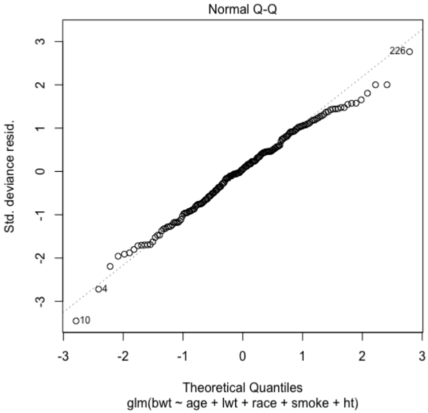


Figure 26: Normal quantile plot of the residuals from the all variables model. There are no important deviations from normality. We can see 3 outliers but they fall relatively closer to the line.

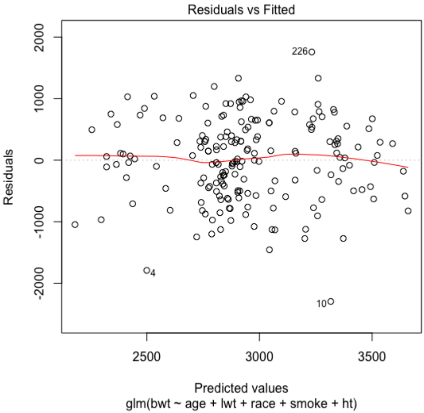


Figure 27: Scatter plot of residual versus predicted values for the birth weight model.

Scatter plot showing distribution of residuals with predicted values. The residuals are distributed fairly uniformly around the residual=0 line, hence the uniformity of variance appears to be satisfied.

- Many studies have yes/no or success/failure response variables. A surgery patient lives or dies; a consumer does or not does not purchase a product after viewing an advertisement.
- Because the response variable in a multiple regression is assumed to have a normal distribution, this methodology is not suitable for predicting such responses. However, there are models that apply the ideas of regression to response variables with only tow possible outcomes.
- One type of model that can be used is called **Logistic Regression**.

[1]

[1] Inputs takes majorly from Introduction to the Practice of Statistics, Fifth edition. David S. Moore. George P. McCabe.